Design and Implementation of OLAP System for Distributed Data Warehouse

Murtadha M. Hamad **Abdullah F. Mahdi**

College of Computer
Anbar University

Received on: 16/10/2012 **Accepted on: 30/01/2013**

الملخص

تشكل عمليات تخزين البيانات والمعالجة التحليلية المباشرة (OLAP) عناصر أساسية لدعم اتخاذ القرارات، التي لديها محل اهتمام متزايد في صناعة قاعدة البيانات. تم تصميم نظام OLAP الموزع الذي يستخدم عدة حاسبات تستند على شبكة محلية. ويمكن إدخال التكنولوجيا الحديثة في مجال توزيع نظام OLAP لحل الاستفسار المعقد والتحليل إلى خدمات مختلفة. في هذا البحث هناك الكثير من المفاهيم النظرية المرتبطة مع OLAP ومستودع البيانات والأنظمة الموزعة للبيانات حيث سيتم تنفيذ العديد من الإجراءات مثل تصميم مكعب البيانات وخوارزمية توزيع وتقسيم مستودع البيانات ونظام دعم القرار (DSS). توضح النتائج العملية آن توزيع البيانات إلى عدة خوادم يمتاز باستفسار ذو سرعة اكبر وينسجم مع التوزيع حيث تم التعامل مع معمارية client -server لتوزيع مستودع البيانات. تم استخدام مفاهيم التحليل الإحصائية في العمل الحالي للحصول على نتائج قابلة للتنبؤ والتي تستخدم للحصول على نتائج مناسبة لدعم القرار.

ABSTRACT

Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. A distributed OLAP system is designed which uses multi microcomputers based local area network. The introduction distributes technology into OLAP system that can disintegrate the complicated query and analysis into different servers. In this paper, there are a lot of theoretical concepts associated with data warehouse and OLAP systems, and distributed data will be the implementation of several measures such as design cubic data and distribution algorithm and division of the data warehouse and decision support system DSS is performed to answer the complicated query. Practical results show that the distribution of data to multiple servers with OLAP system is faster according to the algorithm that has been dealing with client-server architecture to distribute the data warehouse. Statistical analysis concepts are used from current work to get predictable results which can be used to get suitable result DSS.

***Keywords*:** OLAP System, Distribution, Data Warehouse, DSS.

1. INTRODUCTION

On-Line Analytical processing (OLAP) and multidimensional analysis is used for decision support systems to find interesting information from large databases .OLAP stands for Online Analytical Processing. It uses database tables (fact and dimension tables) to enable multidimensional viewing, analysis and querying of large amounts of data. E.g. OLAP technology could provide management with fast answers to complex queries on their operational data or enable them to analyze their company's historical data for trends and patterns [1].

At Present, almost all the OLAP systems are based on only one server. It is excessively expensive to deploy these systems for most middle and small sized enterprises. By introducing distribute technology into OLAP, the system can disintegrate the complicated query and analysis into different servers, then the OLAP application can run on servers with low configuration or microcomputers. Thus, it can make use of available equipment to deploy OLAP system, reduces business cost effectively for companies. There is a profit forecast system for real estate investment that uses OLAP technology to do data analysis. It runs on only one server before, now the defects of OLAP service on the one server begin to emerge with more data and more complex query: such as more response time, slow query speed, slow dependability and scarce expansibility [2].

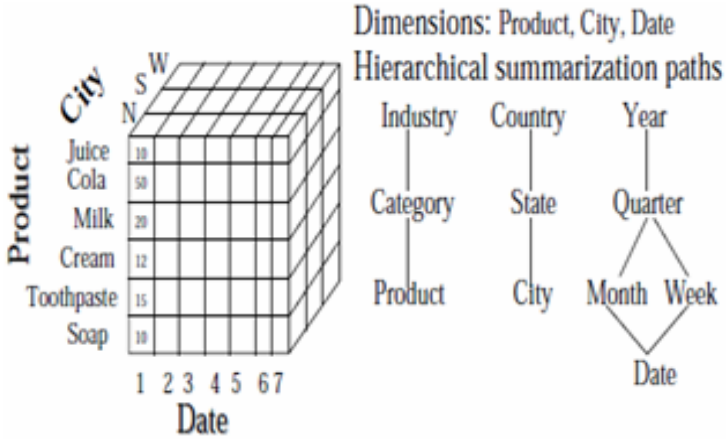

Figure 1. Multidimensional Data

2. Data Cubes and OLAP

Multidimensional systems store data in multidimensional structures which is a natural way to express the multi-dimensionality of the enterprise data and is more suited for analysis. A "cell" in multi-dimensional space represents a tuple, with the attributes of the tuple identifying the location of the tuple in the multi-dimensional space and the *measure* values represent the content of the cell. Data can be organized into a data cube by calculating all possible combinations of GROUP-BYs [2].

A. Basic OLAP Operations

The basic OLAP operations are as follows.

1- Roll up: An operation for moving up the hierarchy level and grouping into larger units along a dimension. Using roll up capability, users can zoom out to see a summarized level of data. Roll up operation is also called the drill up operation.

2- Drill down: An operation for moving down the hierarchy level and stepping down the hierarchy. Using drill down capability, users can navigate to higher levels of detail. Drill down operation is the reverse of roll up operation.

3- Slice: Slicing performs a selection on one dimension of a cube and results in a sub cube. Slicing cuts through the cube so that users can focus on more specific perspectives.

4- Dice: Slicing defines a sub cube by performing a selection on two or more dimensions of a cube[3] and [4].

Proposed algorithm based on the OLAP is as follows:

- **Input:** OLAP query
- **Output:** Values, tables, or graphs of OLAP query result

Step one. User determines the required OLAP query
Step two . Determine the required parameter for analysis
Step three . OLAP Engine determines the required data to OLAP query by assistance of metadata from cube repository
Step four .Select the dimensions and measure from the general table used to design the data cube
Step five .OLAP Engine processes the request and, then displays the result to the user interface.
Step six. The result is presented to the user in the form of values, table, or graphs
End

3. Practical Results

Figure 2 shows the browsing of Reseller Sales Amount measure dimensioned by Country attribute of the Geography user-defined hierarchy in Cube Browser. This is the interest of defining the reference relationships.

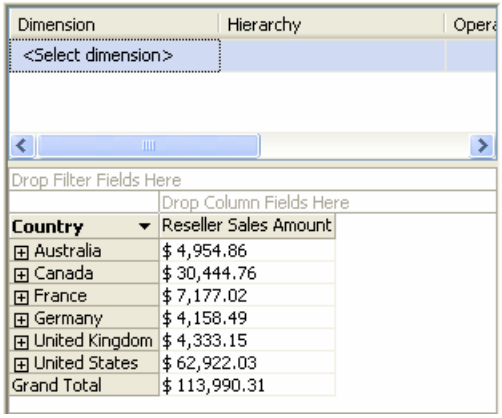

Figure 2. The Reseller Sales Amount Measure Dimensioned by the Country Attribute

4. Distributed Processing

Most networks use distributed processing, in which a task is divided among multiple computer. Instead of a single large machine being responsible for all aspect of a process, separate computers (usual a personal computer or server) handle a subset.

A. Fragmentation Transparency

A unified approach to the distributed design of distributed data warehouse is a data fragmentation, which refers to splitting a huge data set in the fact table into much smaller pieces that can be managed efficiently and minimizing response time to answer a complex query. At this stage, the horizontal fragmentation will be used since they include fragmenting the data and indexed together depending on the proposed framework.

5. OLAP Operation for SALES DW

The multidimensional view allows hierarchies associated with each dimension also to be viewed in a logical manner. Figure 3 explains the hierarchy for the times

dimension and direction of Rull-up and Drill-Down. Show Rull-up operations on times Dimension to aggregating the time Dimension from day to year. In the Drill-down operation to aggregating from year to day.

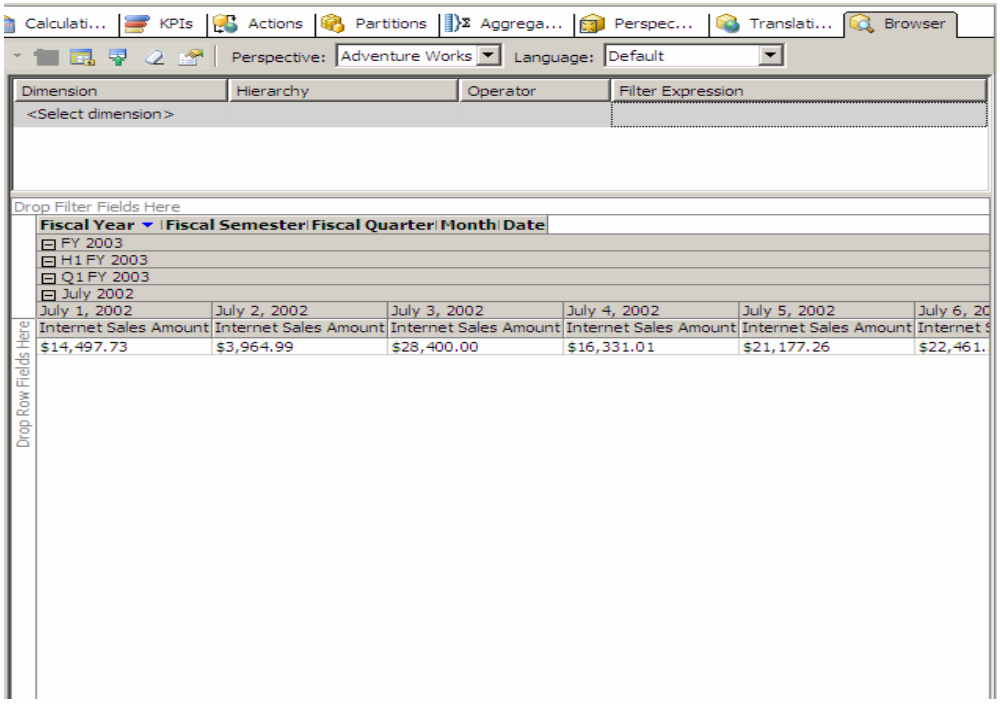

Figure 3. The result for using Reseller using Drill-Down operation OLAP

A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset, for example, if a partition is limited to 2004 data, the partition's data slice should specify the 2004 member of the Time dimension. Figure 4 shows the result of using slice operation of OLAP.

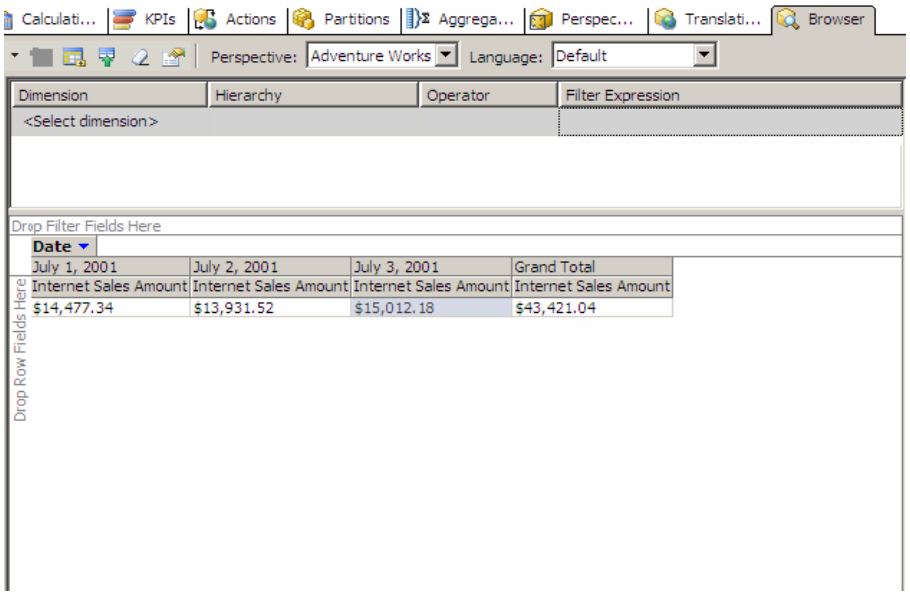

Figure 4. The result of using slice operation of OLAP

6. The Reports of the OLAP Cube Data Sources

A. Resellers Sales Amount

When the report is previewed, the result will look like that in Figure 5 The Figure indicates that the state provinces with large bubble have large reseller sales.

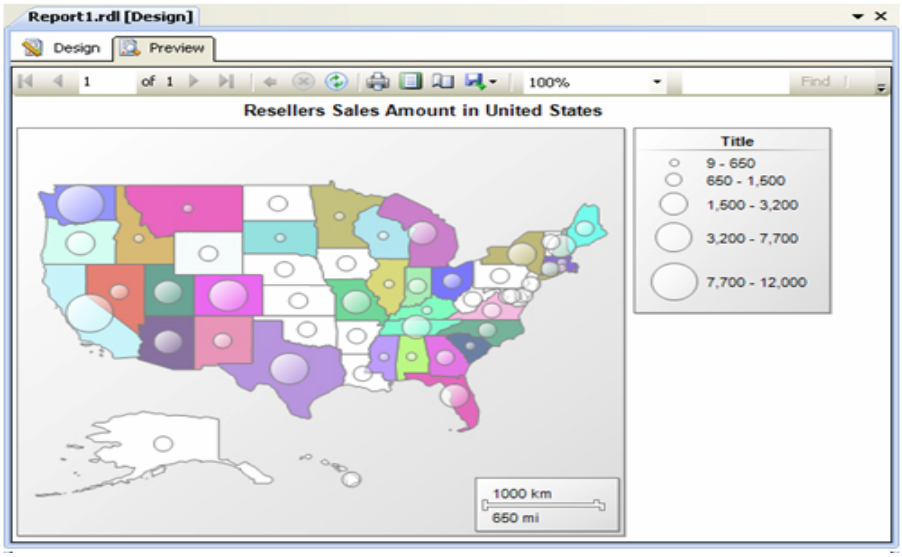

Figure 5. the preview of report

B. Large Reseller Report

When this report is previewed the result will look like that in Figure 6. When you scroll down in the report you can see the Doughnut charts that show business type and product line information of large resellers.

Large Reseller

Reseller	Product Line	Year Opened	Reseller Sales Amount	Indicator
Top Telecom Supply	Accessories	1986	1698.03934947484	★
Retail Mall	Phone	1977	1393.9305376742	★
Rewarding Activities Company	Phone	1974	1385.55781868077	★
Registered Mobiles Store	Accessories	1983	1090.58959378238	★
Citywide Service and Repair	Phone	1987	1007.59251155479	★
Larger Mobiles Shop	Phone	1989	1000.14140630397	★
Metropolitan Equipment	Phone	1986	986.354860103627	★
Sales and Supply Company	Phone	1980	906.488408117444	★
Permanent Finish Products	Accessories	1999	905.162710276341	★
Futuristic Telecom Distributors	Phone	1974	856.227498963731	★
Rally Day Mall	Phone	1987	814.183514853195	★
Friendly Phone Shop	Phone	1980	790.558679015544	★
Outdoor Equipment Store	Phone	1980	780.824961571675	★
Westside Plaza	Applications	1989	745.319906563039	★
Phone Products Distributors	Applications	1978	714.196796240724	★
Odometers and Accessories Company	Applications	1987	700.647264617758	★
Custom Accessories Company	Phone	1998	682.013426639414	★
Grand Discount Store	Accessories	1974	666.353422711571	★

Figure 6. the preview of the report

7. Results and Discussion

In this system the results were compared to the previous results from the centralized DW and through the practical results obtained faster DDW in the query to the decision maker because of the volume of data and partition tables.
the results from several aspects. The most important of which are(Practical, Interplay, Easiness to Implement, Extensible, Performance, Execution time) as follows:

- **Practical:** The algorithmic previous system was working logically and can be used simply in data processing.
- **Easiness to Implement:** Steps of algorithm can be implemented in any structured language.
- **Interplay With the user:** The system interfaces are very clear to the user and are easy to to used and understood.
- **Extensible.** We can develop the project, and other functions can be added to it.
- **Performance:** The system has good performance. This proposed system can be used to develop the DSS.
- **Execution time:** To processing huge amount of data ,we need a few minutes only when we use the proposed system , and wile the speed vs. the high quality will be on it, the data is not affected, especially the processing is running once for all data (See table 1).

Table 1. the response time for Distributed Data Warehouse

Record No	Group By	DDW TIME In(S)
250 000	Yearly	2.129
500000		3.971
750 000		4.891
1000 000		8.785
250 000	Monthly	1.301
500000		2.211
750 000		3.2451
1000 000		5.392
250 000	Daily	1.001
500000		2.692
750 000		3.597
1000 000		4.601

8. Conclusion

1. The OLAP system with one server can be applied simply, but it is inefficient and expensive, so it cannot be applied in most middle and small-sized enterprises.
2. The introduction distributed of a technology into OLAP system can disintegrate the complicated query and analysis into different servers. So, the OLAP application can run on servers with low configuration or microcomputers.
3. Database design depends on the distributed prototype of the foundation, and the number of times, locations and access to application data. Data tables are placed near the application that needs thin more than others.
4. Partition table is useful in distributed database systems, where users and applications with afferent requirements access the same table. Division offers useful and quick access through the local reference of the table.

REFERENCES

[1] Torben Bach Pedersen, Junmin Gu, Arie Shoshani, Christian S. Jensen. Object-extended OLAP querying. Data& Knowledge Engineering, Volume 68, Issue 5, May 2009, Pages 453-480.

[2] De Jun Guan. Data Warehouse And OLAP Technology Applied In Teaching Management [J]. Computer Knowledge and Technology, 2008 6(7):1923 -1924.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Chapter2:Data Warehouse and OLAP Technology for Data Mining, Barnes & Nobles, 2000.

[4] Todd Eavis, George Dimitrov, Ivan Dimitrov, David Cueva, Alex Lopez, Ahmad Taleb. Parallel OLAP with the Sideraserver. Future Generation Computer Systems, Volume 26, Issue 2, February 2010, Pages 259-266.